Speech and Language Processing for Generating Content Description Metadata for Broadcast News

Yoshihiko Hayashi†, Shoichi Matsunaga, and Yoshihiro Matsuo

Abstract

With metadata considered to be a key issue in deploying advanced content services on broadband networks, there is a growing demand for methods of efficiently creating metadata or metadata elements that describe semantic content. This article introduces an approach to improving the efficiency of generating content description metadata for broadcast news programs by using speech recognition and natural language processing.

1. Introduction

The development and popularization of broadband networks is accelerating attempts to distribute and utilize multimedia video content. Advanced content services will use appropriate metadata to describe the content. However, metadata assignment is usually performed manually, and this cost is regarded as being a serious barrier (the metadata bottleneck) to the deployment of such services.

This article introduces our approach to improving the efficiency of generating content description metadata for broadcast news programs by using speech recognition and natural language processing.

2. Content description metadata and its use

Metadata is “data about data”. It provides information about various aspects of contents, from formal information such as media type, data size, and last update date to information related to copyright and terms of use. To make actual content services, we need a method that can efficiently handle these types of information in metadata.

Another important aspect of metadata is how well it describes the content. We call this type of metadata “content description metadata”. Speech extracted from the content and converted into text by speech recognition techniques and the telops[*1] converted into text by character recognition are regarded as metadata and considered as essential information sources for the content description.

If we want to handle video content like news programs, we must decide what kind of information and what format should be included in the content description metadata. Speech and language-based processing can be a powerful tool to capture the semantic intent of a news program, because it is conveyed by the announcer’s speech. Purely audio-visual processing can additionally supply useful information, such as scene change points. However, it cannot extract semantic information.

Most news programs introduce several topics one after another. For keyword searches, the target search unit should be a topic or story. If each story is categorized into a genre like “politics” or “sports”, we can browse the stories in the genre of interest. In addition, if stories have headlines that clearly indicate content, then keyword searching or browsing based on categorization can be made more convenient.

Our system assigns metadata to a program based on this background. The structure of the metadata is schematically shown in Fig. 1. One content (pro-

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: hayashi.yoshihiko@lab.ntt.co.jp

*1 telops: effects produced with a televison opaque projector

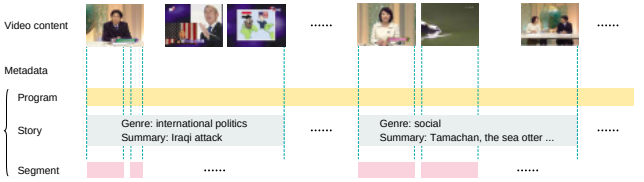

Fig. 1. Structure of content description metadata.

gram) consists of multiple topics (stories), and each story consists of multiple parts (segments). The term segment here means the interval delineated by pauses in the speech. Each interval has an index indicating its start/end point in the content's time frame, so we can easily select/replay any story or segment regardless of the program's length.

3. Utilizing speech recognition and language processing

Figure 2 shows the architecture of our metadata generation (or indexing) system [1]. Actual video programs include not only speech segments but also continuous nonverbal segments. The audio segmentation process segments the input audio stream into intervals each having one of four classification labels: voice, music, noise, and silence. This process is particularly useful in extracting speech segments accurately.

For the speech recognition process, we chose VoiceRex [2], which was developed in our laboratory. In general, to get higher recognition accuracy, we must adapt acoustic/language models to the task at hand. As our system has already learned the words and typical phrases frequently used in news programs, it can recognize news material spoken aloud in a good sound environment with a word error rate of less than 10%.

Speech recognition outputs the words of the segments delineated by speech pauses, and recorded together with links to the program time frame index (vocalization start/end time). In addition, the confidence of the recognition result is recorded as a score, which is used by subsequent language processing functions. Finally, each segment is categorized according to the combination of words in it, and its (estimated) genre is assigned [3].

The speech recognition result is merely a word string, so it needs to be structured for easy searching/access as shown in Fig. 1. We use topic segmentation [4] because it allows us to divide the program into topics (stories). Topic segmentation is based on the idea that when the topic changes, the set of words used changes too. Word meaning is determined from the words used with it. From this viewpoint, this technology constructs, prior to use, a document database (a dictionary) storing the meaning of words as a vector (concept vector: an idea developed by NTT Communication Science Laboratories); it recognizes an obvious change in vector lineage as a topic boundary.

After dividing the program into stories, the system further enriches the metadata being created. It tags

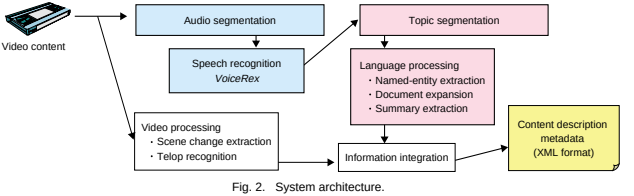

Fig. 2. System architecture.

named entities such as person/place names. If such named entities are well tagged in advance of the search time, a search system can offer advanced search capabilities using these tags. In addition, the system complements the metadata by supplying words retrieved from an external database (*e.g.*, the Internet). This process is effective in decreasing the impact of recognition errors, which are inevitably present in the metadata.

The system also extracts an informative segment of recognized speech from a topic segment as its summary. As described in the next section, this linguistic string is used by the search system to concisely represent the content of a retrieved news story.

The information integration part finally makes decisions on news story segmentation by incorporating the results from the speech-based processes, as well as the ones from the image video-processing tool [5].

In the current system, the structure of content description metadata is based on XML (eXtensible Markup Language). The XML search engine LISTA [6] stores and manages it appropriately.

4. Prototype search system

To examine how content description metadata is used in searching and accessing content, we constructed a prototype search system to access and retrieve news video contents. Its configuration is shown in Fig. 3. The user searches for interesting news stories using a conventional Web browser. When the user selects an interesting news story, the video content corresponding to it is delivered from the distribution server and replayed on the user's terminal.

4.1 Browsing the segmentation result

Figure 4(a) shows an example of a browsing window. The news program is automatically segmented into news stories before the search. At the time of the search, it is represented as a set of news stories as shown in Fig. 4 (a). A news story is represented by a summary, assigned genre, and thumbnails of the story segments. These make it easier for the user to grasp the story's content. As both the summary and genre are based on speech recognition results, they might contain some erroneous strings. However, if the target is a news program, they can provide enough information to give the user a rough understanding of the story. The user can specify/replay a story or segment he is interested in by clicking the displayed thumbnail images or links.

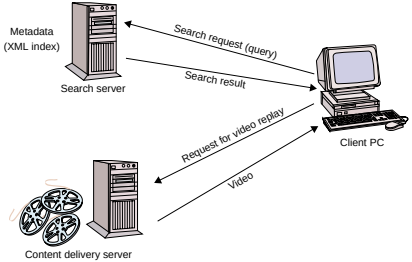

Fig. 3. Search system configuration.

(a) Browsing window (b) Keyword search window (c) Video replay

Fig. 4. Example of search windows.

4.2 Keyword search

The second search option is to input keywords. The user can search using the words in the metadata. The search result is displayed in the same format as a browsing window: a list of thumbnail images, summary, and genre (Fig. 4(b)). The user can replay an interesting story by clicking the links provided by the search results.

Content description metadata includes not only the announcer's speech and the telop words but also the related words complemented by an external database. Therefore, the user can enter words that were not vocalized but are closely related to the story, or words that were vocalized but not recognized. For example, if the word "Mariner's" is seen to complement "Ichiro", the user can locate related stories by using "Mariner's" as a keyword even if its pronunciation is not recognized properly. Thus the use of related words helps offset the degradation of search accuracy caused by speech recognition errors.

In addition, by utilizing the structured search function of the XML search engine, we can set more detailed search conditions. For example, the user can search for keywords that appear only in specific places such as in telops. Similarly, the user can display the news in one genre in a directory-like display by specifying the genre name as a search key.

4.3 Video replay

The user can replay a news video by clicking the links provided by the search result, but content description metadata enables more advanced replay functions. For example, the user can display the texts created by speech recognition in synchronization with the replayed video like captions (Fig. 4(c)).

5. Conclusion

The semantic content of a news program is mostly conveyed by the announcer's speech, and a news program has a clear topic structure. Metadata describing such content can be efficiently generated by the speech and language-based technology described in this article.

In future, we plan to research even more advanced search/access service functions. One idea is to detect and track event topics from incoming multiple news streams. This could lead to an advanced information organization service that provides a user with a timeline summary from a series of related events.

We also intend to improve speech recognition accuracy and provide a more flexible way of adjusting the acoustic/language model. This will enable us to extend the applicability of speech recognition beyond news programs to programs containing speech of various styles.

References

[1] Y. Hayashi, K. Ohtsuki, K. Bessho, O. Mizuno, Y. Matsuo, S. Matsunaga, H. Hayashi, T. Hasegawa, and N. Ikeda, "Speech-based and Video-supported Indexing of Multimedia Broadcast News," in Proc. of SIGIR 2003, (to appear), 2003.

[2] Y. Noda, Y. Yamaguchi, K. Ohtsuki, and A. Imamura, "Speech Recognition Engine VoiceRex," NTT Gijyutsu Journal, Vol. 11, No. 12, pp. 14-17, 1999 (in Japanese).

[3] K. Ohtsuki, T. Matsuoka, S. Matsunaga, and S. Furui, "Topic Extraction based on Continuous Speech Recognition in Broadcast News Speech," IEICE Trans. Vol. E85-D, No. 7, pp. 1138-1144, 2002.

[4] K. Bessho, "Text Segmentation Using Word Conceptual Vectors," Trans. of IPSJ, Vol. 42, No. 11, pp. 2650-2662, 2001 (in Japanese).

[5] Y. Taniguchi, A. Akutsu, and Y. Tonomura, "PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing," in Proc. of ACM Multimedia, pp. 427-436, 1997.

[6] J. Tomita, "XML Document Search System LISTA," NTT R&D, Vol. 52, No. 2, pp. 85-91, 2003 (in Japanese).

Yoshihiko Hayashi
Senior research engineer, supervisor, NTT Cyberspace Laboratories.
He received the B.S., M.S., and Dr.Eng. degrees from Waseda University in 1981, 1983, and 1996, respectively. In 1983, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation (now NTT), Japan. He was a visiting researcher at CSLI, Stanford University from 1994 to 1995. His current interests include natural language processing, cross-language/cross-media information access.

Shoichi Matsunaga
Senior research engineer, NTT Cyberspace Laboratories.
He received the Dr.Eng. degree from Kyushu University, Fukuoka, Japan in 1992. Since joining NTT, he has been working on automatic speech recognition. From 1994 to 1996, he was a researcher at ATR Interpreting Telecommunications Research Laboratories. He is the member of the Acoustical Society of Japan, IEICE (Institute of Electronics, Information and Communication Engineers), and IEEE.

Yoshihiro Matsuo
Senior research engineer, NTT Cyberspace Laboratories.
He received the B.S. and M.S. degrees from Osaka University in 1988 and 1990, respectively. In 1990, he joined Communications and Information Processing Laboratories, NTT, Japan. His current interests include machine translation and cross-language/cross-media information access.